STATISTICAL RESEARCH REPORT No 3
Institute of Mathematics 1979
University of Oslo

THE BEHRENS-FISHER PROBLEM

by

Grete Usterud Fenstad

1. SUMMARY

Several statistical methods have been recommended for use in the Behrens-Fisher situation. We have compared the significance levels and the power functions of some of them by analytical methods and stochastic simulations.

If the sample sizes are unequal it is well known that the asymptotic significance level of the T-test can be far from the nominal level if the sample variances are not equal. We have shown that the significance level can be even worse for finite samples.

The Wilcoxon test also fails for unequal sample variances. However, contrary to the T-test it behaves better for finite samples than for infinite samples.

One of the tests introduced by Welch (1937) and which later seems to have been ignored, appears to have significance level closer to the nominal level.

Some adaptive tests which also were examined have significance levels even closer to the nominal level.

For some parameter values the power functions of the tests have been estimated and also in this respect the adaptive tests do better than the others.

The power function of the exact test introduced by Scheffé (1943) seems to compare well with the non-adaptive tests.

2. INTRODUCTION

Testing the hypothesis of equality of the means of two normal distributions with unknown variances is called the Behrens-Fisher problem. More precisely we have the following situation:

Let

(1) $X_1,\ldots,X_m$ i.i.d. $N(\xi,\sigma^2)$ and independent of
$Y_1,\ldots,Y_n$ i.i.d. $N(\eta,\tau^2)$.

The problem is that of testing the hypothesis

(2) $$H : \xi = \eta$$

against all possible alternatives without assuming equality of the variances σ^2 and τ^2.

We shall nevertheless first consider the special case where $\sigma^2 = \tau^2$ in (1). The likelihood ratio test with significance level ε rejects the hypothesis when $|T| > t_{\varepsilon/2;m+n-2}$, where

$$(3) \qquad T = (\bar{X}-\bar{Y}) \Big/ \sqrt{(\frac{1}{m}+\frac{1}{n})\frac{Z_1+Z_2}{m+n-2}}$$

and $t_{\alpha;f}$ is the upper α-fractile in the t-distribution. Here and in the following se use the notation:

$$\bar{X} = (1/m)\sum_{i=1}^{m} X_i \ , \ \bar{Y} = (1/n)\sum_{i=1}^{n} Y_i \ , \ Z_1 = \sum_{i=1}^{m}(X_i-\bar{X})^2, \ Z_2 = \sum_{i=1}^{n}(Y_i-\bar{Y})^2.$$

It is well known that this test (the T-test) is UMP among unbiased ε-tests.

The statistic T is under the hypothesis t-distributed with $m+n-2$ degrees of freedom, and thus the significance level is independent of $\sigma^2(=\tau^2)$.

For large m and n T has approximately a normal distribution with unit variance. This is the case also under more general assumptions, namely

$X_1,\ldots,X_m$ i.i.d. with expectation ξ and variance σ^2 and

$Y_1,\ldots,Y_n$ i.i.d. with expectation η and variance τ^2.

It is, however, necessary that the variances are the same in both samples.

We shall not here consider robustness of the T-test under deviations from normality, but rather robustness of the T-test (and other tests) under deviations from equal variances, that is the behaviour of T under the assumptions (1). We shall mainly concentrate on the significance levels, as the power of the tests turns out to depend very much on the significance level.

3. ASYMPTOTIC SIGNIFICANCE LEVELS OF THE T-TEST AND THE W-TEST

In Chapter 2 some of the properties of the T-test in the special case of equal variances were summarized. The T-test, however, is also much in use in situations where there is no reason to assume equal variances.

One way to illustrate the behaviour of the T-test for $\sigma^2 \neq \tau^2$ or $\theta = \sigma^2/\tau^2 \neq 1$ is to study its asymptotic significance level as $m,n \to \infty$. Assuming $m/n \to \nu$, the asymptotic distribution of $\sqrt{(\nu\theta+1)/(\nu+\theta)} \cdot T$ is $N(0,1)$ under the hypothesis, and hence the asymptotic significance level is

$$\varepsilon(\theta) = 2[1 - \Phi(\sqrt{(\nu\theta+1)/(\nu+\theta)}\, u_{\varepsilon/2})],$$

where Φ is the cumulative standard normal distribution and $u_{\varepsilon/2}$ its upper $\varepsilon/2$-fractile.

It is seen that $\varepsilon(\theta)$ depends on θ. For $\nu < 1$ $\varepsilon(\theta)$ is increasing with θ and for $\nu > 1$ $\varepsilon(\theta)$ is decreasing with θ. For $\nu = 1$ one obtains the promising result that $\varepsilon(\theta) \equiv \varepsilon$ for all θ. Therefore it is often recommended that m and n should be (almost) equal.

The asymptotic significance level $\varepsilon(\theta)$ has been tabulated for the nominal level $\varepsilon = 0.05$ and for some $\nu \leq 1$ in Table 1. The values of $\varepsilon(\theta)$ for $\nu > 1$ are obtained from the same table by entering the table with $1/\nu$ and $1/\theta$.

ν \ θ	0	1/8	1/4	1/2	1	2	4	8	∞
0	0	-	-	0.006	0.05	0.166	0.327	0.488	1
1/10	-	-	0.001	0.010	0.05	0.138	0.252	0.356	0.535
1/3	0.001	0.003	0.008	0.020	0.05	0.098	0.150	0.194	0.258
1/2	0.006	0.011	0.016	0.028	0.05	0.080	0.110	0.133	0.166
1	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05

TABLE 1

The asymptotic significance level of the T-test with significance level $\varepsilon = 0.05$, $\theta = \sigma^2/\tau^2$ and $m/n \to \nu$ (- means $< 10^{-3}$).

In Fig. 1 $\varepsilon(\theta)$ has been drawn for $\nu = 1$, 1/2 and 1/3. The most striking fact is how fast $\varepsilon(\theta)$ changes with θ in the neighborhood of $\theta = 1$ when $\nu \neq 1$. One thus get a warning from the asymptotic calculations to be careful using a T-test if $m \neq n$ and one is not quite sure that $\theta = 1$. It will in fact later be demonstrated that the limit $\varepsilon(\theta)$ as $m,n \to \infty$ is obtained from above, such that the situation is even worse for finite m and n than the asymptotic significance levels show.

A competitor to the T-test in case of equal variances is the Wilcoxon or Mann-Whitney test, here called the W-test. We find it most convenient to work with the Mann-Whitney statistic,

$$W = \text{number of pairs } (X_i, Y_j) \text{ with } X_i < Y_j.$$

The hypothesis is rejected if $|W_{XY} - m \cdot n/2| > w_{\varepsilon/2;m,n}$ where $w_{\varepsilon/2;m,n}$ is determined such that the significance level is (approximately) ε in case of equal variances, $\theta = 1$.

To obtain the asymptotic significance level of this test we use the fact that the distribution of $(W-EW)/\sqrt{\operatorname{var} W}$ tends to a standard normal distribution as $m,n \to \infty$, see Lehmann (1975) pp. 69-70. Under the assumptions (1) and (2) one obtains

$$E\,W = m\,n/2 \quad \text{and}$$

$$\operatorname{var} W = m\,n\,[\pi/2+(n-1)\operatorname{Arcsin}(\theta/(\theta+1))+(m-1)\operatorname{Arcsin}(1/(\theta+1))]/2\pi.$$

Further $w_{\varepsilon/2;m,n} \approx u_{\varepsilon/2}\sqrt{mn(m+n+1)/12}$ for large m and n, which leads to a significance level of the W-test which is approximately equal to

$$(4) \qquad 2\left[1-\Phi\left(u_{\varepsilon/2} \cdot \sqrt{mn(m+n+1)/12}\ \Big/ \sqrt{\operatorname{var} W}\right)\right]$$

and an asymptotic significance level

$$\varepsilon_W(\theta) = 2[1-\Phi(u_{\varepsilon/2} \cdot k(\theta))]$$

where $k(\theta) = \sqrt{(\nu+1)\pi/6}\ \Big/ \sqrt{\operatorname{Arcsin}(\theta/(\theta+1))+\nu\operatorname{Arcsin}(1/(\theta+1))}$.

Similar to $\varepsilon(\theta)$ of the T-test, $\varepsilon_W(\theta)$ depends on θ. However, $\varepsilon_W(\theta)$ is not a monotone function of θ, it is also not independent of θ for $\nu = 1$.

The asymptotic significance level $\varepsilon_W(\theta)$ has been tabulated for the nominal level $\varepsilon = 0.05$ and for some $\nu \leq 1$ in Table 2. The values of $\varepsilon_W(\theta)$ for $\nu > 1$ are obtained from the same table by entering the table with $1/\nu$ and $1/\theta$.

ν \ θ	0	1/8	1/4	1/2	1	2	4	8	∞
0	0	-	0.002	0.015	0.050	0.097	0.141	0.175	0.258
1/10	-	0.002	0.006	0.021	0.050	0.089	0.126	0.157	0.235
1/3	0.024	0.018	0.022	0.032	0.050	0.074	0.101	0.124	0.191
1/2	0.050	0.032	0.033	0.039	0.050	0.067	0.087	0.105	0.166
1	0.110	0.068	0.059	0.052	0.050	0.052	0.059	0.068	0.110

TABLE 2

The asymptotic significance level of the W-test with significance level $\varepsilon = 0.05$, $\theta = \sigma^2/\tau^2$ and $m/n \to \infty$ (- means 10^{-3}).

In Fig. 1 $\varepsilon_W(\theta)$ has been drawn for $\nu = 1$, 1/2 and 1/3. Even if $\varepsilon_W(\theta)$ behaves somewhat better than $\varepsilon(\theta)$ of the T-test, it still changes too fast with θ in the neighborhood of $\theta = 1$. It can, however, be derived from (4) that for large m and n the limit $\varepsilon_W(\theta)$ for those θ for which $\varepsilon_W(\theta) > 1$ is obtained from below, such that - contrary to the T-test - the situation is better for finite m and n than the asymptotic significance levels show.

4. ALTERNATIVE TESTS

Neither the T-test nor the W-test is designed to take care of the situation with θ completely unknown, so one should start anew and look for other tests.

Since the family of distributions is exponential any power function is continuous. Hence any unbiased test is also similar. However, the set of sufficient statistics under the hypothesis, $\bar{X}$, $\bar{Y}$, Z_1, Z_2, is not boundedly complete so the usual theory (Lehmann (1959) Ch.4.4) cannot be applied.

Reducing by sufficiency and invariance leads to consider the statistics $(\bar{X}-\bar{Y})/\sqrt{Z_1+Z_2}$ and Z_1/Z_2. Linnik (1963) proved that under reasonable restrictions the only similar test based on these statistics is the trivial one which rejects with probability ε.

Linnik's result doesn't mean that there exists no non-trivial test, with significance level independent of θ. Scheffé (1943) introduced the following test:

Assuming $m \leq n$, set

$$S_i = X_i - \sqrt{\frac{m}{n}}\, Y_i + \frac{1}{\sqrt{mn}} \sum_{j=1}^{m} Y_j - \bar{Y} \quad , \quad i=1,\ldots,m$$

which are independent, identically normally distributed $N(\xi-\eta, \sigma^2 + \frac{m}{n}\tau^2)$. Since

$$(5) \qquad S = \bar{S} \cdot \sqrt{m} \Big/ \sqrt{\frac{1}{m-1} \sum_{i=1}^{m} (S_i-\bar{S})^2}$$

under the hypothesis is t-distributed with $m-1$ degrees of freedom, the test which rejects when $|S| > t_{\varepsilon/2;m-1}$ is an unbiased test with significance level ε independent of θ.

The power function of the S-test can be expressed by the t-disribution with $m-1$ degrees of freedom and non-centrality parameter $\delta = (\xi-\eta)/\sqrt{\sigma^2/m + \tau^2/n}$ or by the corresponding non-central F-distribution.

Notice that the numerator of S is $\bar{S} = \bar{X} - \bar{Y}$ and that $\sum_{i=1}^{m} (S_i-\bar{S})^2/m(m-1)$ is an unbiased estimator of its variance $\sigma^2/m + \tau^2/n$. The estimator fails, however, to utilize the observations $Y_{m+1},\ldots,Y_n$, such that the statistic S may change if the Y's are permuted. Scheffé (1970) claims that he suggested the S-test before he had much consulting experience, and "since then I have never recommended the solution in practice".

It is still natural to base a test of (2) on $\bar{X} - \bar{Y}$ compared with an estimator of its standard deviation based on sufficient statistics. The test which rejects when $|U| > u_o$, where

$$(6) \qquad U = (\bar{X}-\bar{Y}) \Big/ \sqrt{Z_1/m(m-1) + Z_2/n(n-1)}$$

and u_o is a constant, had often been employed when Welch (1937) compared some of its properties with the properties of the T-test.

The test which rejects when $|V|>v_o$, where

$$(7) \qquad V = (\bar{X}-\bar{Y})\Big/\sqrt{Z_1/m(m-3) + Z_2/n(n-3)}$$

and v_o is a constant, was suggested in this same paper by Welch, but have not been studied since then. The reasoning behind this test is found in Section 5.

The asymptotic distributions of U and of V under the hypothesis are both N(0,1) and hence the asymptotic significance levels are independent of θ for any ν.

5. DISTRIBUTIONS OF THE TEST STATISTICS FOR FINITE m AND n

In the two previous chapters we have considered the asymptotic distributions of the T-, U- and V-statistics under the hypothesis, and we shall now turn to their distributions for finite m and n.

They may all be written in the form

$$(8) \qquad R\Big/\sqrt{\alpha\cdot W_1/f_1 + \beta\cdot W_2/f_2}$$

where $R = (\bar{X}-\bar{Y})\Big/\sqrt{(\sigma^2/m)+(\tau^2/n)}$, $W_1 = Z_1/\sigma^2$, $W_2 = Z_2/\tau^2$, $f_1 = m-1$ and $f_2 = n-1$. R, W_1 and W_2 are independent, R has a standard normal distribution, $W_1(W_2)$ is χ^2-distributed with $f_1(f_2)$ degrees of freedom. The parameters α and β depend on m,n,θ and on the actual statistic. From these facts the distribution of the statistic (8) may be derived. An expression for the density of the statistic is found in Sverdrup (1964) Chapter XIV.3 (10). Calculation of any probability by use of this expression leads to numerical integration.

We have instead chosen to follow Welch (1937) who derives an approximate distribution of the statistic (8). He uses the fact that

$$(9) \qquad \alpha \cdot W_1/f_1 + \beta \cdot W_2/f_2$$

is approximately distributed as $c \cdot W/f$ where W is χ^2-distributed with f degrees of freedom. The constants c and f are determined such that the two first moments are equal, i.e.

$\alpha + \beta = c$ and $\alpha^2/f_1 + \beta^2/f_2 = c^2/f$ or

$$(10) \qquad c = \alpha + \beta \quad , \quad f = \frac{(\alpha+\beta)^2}{\alpha^2/f_1+\beta^2/f_2} \ .$$

This approximation is good also for small f_1 and f_2. Applying this to the statistic (8) gives that it is approximately distributed as

$$(11) \qquad R/\sqrt{c \cdot W/f} = T_f/\sqrt{c}$$

where T_f is t-distributed with f degrees of freedom. Thus a common approximate expression for the significance levels of the three tests is

$$(12) \qquad \varepsilon(\theta) = 2[1 - T_f(\sqrt{c} \cdot k)]$$

where k is the critical value of the test and $T_f(\cdot)$ is the cumulative t-distribution with f degrees of freedom. The dependence of $\varepsilon(\theta)$ on m and n will be taken for granted and will not be marked in the notation.

To utilize (12), c and f for the three tests are needed. After some calculation one obtains - with obvious notation:

$$(13) \quad \begin{cases} c_T = \dfrac{1/m+1/n}{\theta/m+1/n} \cdot \dfrac{(m-1)\theta+(n-1)}{(m-1)+(n-1)} & , \ f_T = \dfrac{[(m-1)\theta+(n-1)]^2}{(m-1)\theta^2+(n-1)} \\ c_U = 1 & , \ f_U = \dfrac{(\theta/m+1/n)^2}{\theta^2/m^2(m-1)+1/n^2(n-1)} \\ c_V = \dfrac{(m-1)\theta/m(m-3)+(n-1)/n(n-3)}{\theta/m+1/n} & , \ f_V = \dfrac{[(m-1)\theta/m(m-3)+(n-1)/n(n-3)]^2}{(m-1)\theta^2/m^2(m-3)^2+(n-1)/n^2(n-3)^2} \end{cases}$$

An approximate expression for the power functions of these tests can be derived in a similar way, noticing that R is normally distributed with expectation $\delta = (\xi-\eta)/\sqrt{(\sigma^2/m)+(\tau^2/n)}$ and variance 1. The statistic (8) times $\sqrt{c}$ is then non-central t-distributed with f degreees of freedom and non-centrality parameter δ, and the power function is

$$(14) \qquad \beta(\delta;\theta) = 1 - T_f(\sqrt{c}\,k;\delta) + T_f(-\sqrt{c}\,k;\delta)$$

where $T_f(\cdot;\delta)$ is the cumulative t-distribution with f degrees of freedom and non-centrality parameter δ.

We are now also able to explain Welch's choice of the statistic V. Any test statistic of the form $(\bar{X}-\bar{Y})/\sqrt{aZ_1+bZ_2}$ is approximately distributed as (11) with $\alpha = a(m-1)/(\theta/m+1/n)$ and $\beta = b(n-1)/(\theta/m+1/n)$. The variance, $\gamma(\theta)$, of (11) under the hypothesis is

$$\gamma(\theta) = \frac{1}{c}\cdot\frac{f}{f-2} = \frac{\alpha+\beta}{(\alpha+\beta)^2 - 2[\alpha^2/(m-1)+\beta^2/(n-1)]}$$

according to (10). Welch's idea was to choose a and b such that this variance is as stable as possible when θ goes from 0 to ∞. He required $\gamma(0) = \gamma(\infty)$, which implies $a = [m(m-3)]^{-1}$ and $b = [n(n-3)]^{-1}$ or a multiple thereof.

6. CHOICE OF CRITICAL VALUES

To compare the properties of the tests one has to fix the critical values.

The critical values of the T-test and the W-test have been chosen such that their respective significance levels are ε when $\theta = 1$.

For the S-test one obviously chooses the critical value which gives significance level ε for all θ.

It remains to fix u_o and v_o for the U- and V-test. We first look to the U-test which has an approximate significance level $\varepsilon_U(\theta) = 2[1-T_{f_U}(u_0)]$ where f_U is given by (13). Studying f_U as a function of θ gives

$$\max_\theta f_U = m+n-2 \quad \text{for} \quad \theta = m(m-1)/n(n-1)$$

$$\min_\theta f_U = \begin{cases} m-1 & \text{for } \theta = \infty \text{ if } m \le n \\ n-1 & \text{for } \theta = 0 \text{ if } m > n\,. \end{cases}$$

Since $\varepsilon_U(\theta)$ decreases with increasing f_U, we obtain

$$\varepsilon_U(m(m-1)/n(n-1)) \le \varepsilon_U(\theta) \le \max\{\varepsilon_U(0), \varepsilon_U(\infty)\}\,.$$

Thus choosing

$$u_0 = \begin{cases} t_{\varepsilon/2}\,;\, m-1 & m \le n \\ t_{\varepsilon/2}\,;\, n-1 & m > n \end{cases}$$

we obtain $\varepsilon_U(\theta) \le \varepsilon$ for all θ. This will always be done in the following examples.

The same simple argument cannot be applied to $\varepsilon_V(\theta) = 2[1-T_{f_V}(\sqrt{c_V}\cdot v_0)]$ as also c_V depends on θ. Since U and V have the same limiting distribution, $\varepsilon_V(\theta)$ will obtain its maximum for θ close to 0 or ∞ for large m and n. For moderate m and n we have also chosen v_0 such that $\max\{\varepsilon_V(0),\varepsilon_V(\infty)\} = \varepsilon$. (This is seen to work well in the examples of Section 7.) Since

$$\varepsilon_V(0) = 2[1-T_{n-1}(\sqrt{(n-1)/(n-3)}\,v_0)] \text{ and } \varepsilon_V(\infty) = 2[1-T_{m-1}(\sqrt{(m-1)/(m-3)}\,v_0)]$$

we choose v_0 equal to the larger of $\sqrt{(n-3)/(n-1)}\,t_{\varepsilon/2;n-1}$ and $\sqrt{(m-3)/(m-1)}\,t_{\varepsilon/2;m-1}$. Unfortunately $\sqrt{(m-3)/(m-1)}\,t_{\varepsilon/2;m-1}$ is not a monotone function of m (see Table 3), so one should be a bit careful in the choice of v_0.

Notice that with this choice of v_0 the U- and V-tests are identical for $m = n$.

ε \ m	4	6	8	10	20	50	∞
0.01	3.372	3.123	2.958	2.866	2.706	2.625	2.576
0.05	1.837	1.991	1.998	1.995	1.980	1.968	1.960
0.10	1.359	1.561	1.601	1.617	1.636	1.642	1.645

Table 3

The expression $\sqrt{(m-3)/(m-1)}\,t_{\varepsilon/2;m-1}$ is tabulated for $\varepsilon = 0.01,\ 0.05,\ 0.10$ and some values of m.

7. NUMERICAL EXAMPLES

We now apply (12) and (13) from Section 5 with critical values as described in Section 6 to calculate the approximate significance levels of the T-, U- and V-tests.

This has been done for the T-test for $m = n = 6$ and 10 and $m = 5$, $n = 15$. In Fig. 2 these significance levels can be compared to the asymptotic ones for $\nu = 1$ and $\nu = 1/3$. The figure

suggests that the asymptotic levels are obtained from above. It also shows how the significance level fails for $m = 5$, $n = 15$ even if one restricts θ to $1/4 < \theta < 4$, say. We find that in this case the T-test should not be used if it is any doubt about $\theta = 1$. If $\theta > 1$ the significance level is too high, if $\theta < 1$ the power function is too low as will soon be seen.

In Fig. 3 the significance levels of the T-, U- and V-tests have been compared for $m = n = 10$ (Fig. 3(a)) and $m = 5$, $n = 15$ (Fig.3(b)). It is specially seen that the significance level of the V-test vary less about ε than the significance level of the U-test when $m = 5$, $n = 15$.

We shall also not forget the S-test which has significance level ε for all m, n and θ.

In (14) (Section 5) we gave an approximate expression for the power functions of the T-, U- and V-tests. Evaluating the power functions thus involves the values of cumulative non-central t-distributions (or F-distributions) with number of degrees of freedom which usually are not an integer. Since the formula (14) already is an approximation, we felt the necessary interpolation on the degrees of freedom made the results rather inaccurate. We therefore performed stochastic simulation to estimate the power functions in Fig. 4. Each point is based on 3000 samples with $m = 5$, $n = 15$ giving a standard deviation of at most $0.5 / \sqrt{3000} = 0.009$.

In Fig. 4(a) we show the power functions for $\theta = 1$. In this case the T-test is known to be uniformly most powerful. For $\delta = 3$ the difference between the power of the T-test and the power of the V-test is about 17%. The power function of the W-test was also estimated and found to be slightly lower than the power of the T-test, nowhere is the difference more than 3%. It is also interesting to compare with the exact S-test. Since its power function is the same for all θ and also in order not to overload the figures, its power function is given in Table 4. It is seen to behave slightly better than V for $\delta = 1$, but behaves worse for $\delta = 3, 4$ and 5. To get another idea of the accuracy of the estimated power functions, the estimated power of the T-test can be compared to the exact power function given in Table 4.

δ	0	1	2	3	4	5
Power of S-test for all θ	0.050	0.122	0.336	0.619	0.843	0.955
Power of T-test for $\theta = 1$	0.050	0.157	0.473	0.810	0.966	0.997

Table 4

The powers of the S-test and the T-test as a function of $\delta = (\xi-\eta)/\sqrt{\sigma^2/m + \tau^2/n}$ for $m = 5$ and $n = 15$. For the T-test $\theta = \sigma^2/\tau^2 = 1$ has been assumed.

In Fig. 4 (b) and (c) we show the power functions for $\theta = 4$ and $\theta = 1/4$ respectively. The power function of the V-test for $\theta = 4$ has been deleted since it is almost coinciding with the power function of the U-test, the difference is everywhere less than 1%.

However, the power of the T-test is always higher than the power of the U-test. The high values of the power of the T-test for $\theta = 4$ is uninteresting because of also the high significance level of about 16% (see Fig 2). For $\theta = 1/4$ it is seen that the V-test is to be preferred. It should also be noted that among these three tests the one with highest significance level also has largest power. This should indicate the use of the V-test, see Fig. 3.
Comparing with the power of the S-test given in Table 4 the S-test is a little more powerful than the V-test for small δ and less powerful for larger δ. This is most pronounced for $\theta = 1/4$.

8. ADAPTIVE TESTS

The tests considered in the previous sections have all been tests with fixed critical values. Approximate distributions of the test statistics T, U and V were derived in Section 5. The test statistics were all distributed as $T_f/\sqrt{c}$, where f and c depend on θ. The fixed critical values of the U- and V-tests were chosen such that the significance levels should never exceed a chosen value ε.

Another method which probably would level out the significance level is to estimate θ by $\hat{\theta} = (Z_1/(m-1))/(Z_2/(n-1))$ - or some other estimator. One then obtains estimators, $\hat{f}$ and $\hat{c}$, of f and c by replacing θ with $\hat{\theta}$. Finally, reject the hypothesis if the test statistic is greater than $t_{\varepsilon/2;\hat{f}}/\sqrt{\hat{c}}$.

This idea was applied to the U-test by Welch (1936). The exact probability of rejecting by the adaptive U-test has later been derived by Wang (1971). The probabilities of rejecting by the adaptive T- and V-tests may be derived in a similar way. To compute the probabilities of the adaptive U-test Wang performed numerical integration. We have chosen to estimate the significance levels and powers of the three adaptive tests by stochastic simulation. Each point is based on 3000 samples giving a standard deviation of about $\sqrt{0.05 \cdot 0.95/3000} = 0.004$ for a nominal level of 5% and a standard deviation of at most 0.009 for the power functions. The estimated number of degrees of freedom $\hat{f}$, is usually not an integer. To determine the fractiles $t_{\varepsilon/2;\hat{f}}$ we used the following approximation due to Wang (1971)

$$t^2_{\alpha;f} = f \cdot \left[e^{u_\alpha^2 / g(f)} - 1 \right]$$

where u_α is the upper α-fractile of the normal distribution and

$$g(f) = 0.9990\ f - 0.480\ .$$

For $\alpha = 0.025$ and $8 \leq f \leq 18$ the difference between the exact fractile and this approximation is less than 0.0005.

It turns out that for $m = n$, the three tests are equivalent. In Fig. 5 (a) we show the common estimated significance level for the three tests for $m = n = 10$ and ε equal to 5%. We have only performed simulations for $\theta \geq 1$, as the significance level is symmetric about $\theta = 1$ in this case. Within the two dotted lines one finds for each θ the acceptance region for a 5%-test of the hypothesis that the significance level is equal to 5% based on 3000 samples.

In Fig. 5 (b) we show the significance levels of the three tests for $m = 5$, $n = 15$ and ε equal to 5%. Among these three tests the adaptive test based on V performs best, and again one cannot rely on the significance level of the test based on T. Comparing with Fig. 3 the significance levels of the adaptive tests are more stable about the nominal significance level of 5%.

In Fig. 6 we show the estimated power functions for the three tests. As for the tests with fixed critical values we obtain that

the test with largest significance level also has the largest power. Of visual reasons we have deleted the adaptive U-test in Fig. 6 (c), it is for each δ estimated to lie between the adaptive T- and V-tests. The power function of the S-test falls below all the power functions drawn in Fig. 6 , the greatest difference is found for $\theta = 1/4$.

9. OTHER TESTS

There have been suggested other tests for the Behrens-Fisher problem than those considered here.

A Bayesian approach leads to reject the hypothesis if $|U| > d(\hat{\omega})$ where $d(\hat{\omega})$ is the upper $\varepsilon/2$ fractile of the Behrens distribution with $m-1$ and $n-1$ degrees of freedom and angle $\hat{\omega}$ depending on Z_1/Z_2 . These results may be found in Lindley (1965) and the fractiles are tabulated in Fisher and Yates (1963). Comparing $d(\hat{\omega})$ with $t_{\varepsilon/2;\hat{f}}$ for the adaptive U-test, one sees that the significance level of the Bayesian test always is lower than for the adaptive U-test.

Welch (1947) tries to determine a function $h(Z_1,Z_2;\varepsilon)$ such that

$$P(\bar{X}-\bar{Y} > h(Z_1,Z_2,\varepsilon)) = \varepsilon \ .$$

He obtains an approximate solution which has been tabulated by Aspin (1948) and (1949). Welch shows, however, that this test is approximated by the adaptive U-test. Their significance levels have been compared by Wang (1971), who concludes that it seems reasonable to use the adaptive U-test because it is less tedious to compute its critical values.

A simulation study of the adaptive test based on the studentized Wilcoxon statistic was recently performed by Yuen Fung (1979), who found that the power of this test did not perform better than the adaptive U-test in the situation we consider.

Finally one could first perform a test of the hypothesis $\theta = 1$. If this hypothesis is rejected, perform a V-test, and if it is not rejected, perform a T-test. One would hopefully obtain a test with the nice properties of the power of the T-test for $\theta = 1$ and get rid of the bad behaviour of the significance level of the T-test for $\theta \neq 1$. To this aim the test of the hypothesis of $\theta = 1$ must have a large significance level. We performed simulations also in this direction for $m = 5$, $n = 15$ and found that a significance level of 50% was not large enough to get rid of the bad performance of the significance level of the T-test in the neighbourhood of $\theta = 1$.

10. CONCLUDING REMARKS

We have through some numerical examples studied the behaviour of both significance level and power function of some tests in use in the Behrens-Fisher situation. It has been necessary to restrict oneself to only a few values of m and n, nevertheless it seems possible to draw some conclusions.

When $m \neq n$, one has to weigh the optimality of the T-test for $\theta = 1$ against its real bad properties for $\theta \neq 1$. If $m > n$ and $\theta > 1$, the significance level of the T-test may be much larger than the nominal level. If $\theta < 1$ the significance level is lower than the nominal level which leads to low power function. If $m \neq n$ one should avoid the T-test unless being convinced of $\theta = 1$.

From the connection between the significance level and power function one should aim at a test with significance level as close to the nominal one as possible. Of the tests with fixed critical value the V-test is the best one in this respect.

Then turning to the adaptive tests in Section 8 one obtains the significance levels even more straightened out. Again the test based on V comes out most favorable. The power functions of the adaptive tests (see Fig. 6) also have improved compared to the tests with fixed critical values (see Fig. 4).

Unfortunately the adaptive V-test does not compare as favorable with the T-test with fixed critical value as could be wished for $\theta = 1$. Comparing Fig. 4 (a) and Fig. 6 (a) one finds the largest difference between the estimated powers of the (optimal) T-test and the adaptive V-test to be 12% for $\delta = 3$.

For $\theta = 1$ also the W-test (Wilcoxon) performs better than the U-test and V-test and also than their corresponding adaptive tests. Similar to the T-test, however, it doesn't work well for $\theta \neq 1$.

Finally the S-test, which has significance level equal to the nominal level for all θ, has even lower power than the adaptive U-test and V-test for $\theta = 1$. For those θ we have estimated the power functions the S-test is comparable to the U-test and V-test with fixed critical values, while its power is lower than for the adaptive tests. Usually the S-test is ruled out because the statistic might take different values after renumbering the observations.

As a general advice one should use the V-test if $\theta \neq 1$, either one considers adaptive or non-adaptive tests. The adaptive V-test increases the power function of the V-test with fixed critical value with 5 - 10 % . On the other side this gain has to be weighed against the more extensive work in calculating the critical value for the adaptive test than for the test with fixed critical value.

Acknowledgements

I wish to thank Lars Walløe for stimulating discussions and for helpful comments. I also wish to thank Morten Kjærnes for assistance during the simulations and Torhild Isaksen for drawing the figures.

REFERENCES

Aspin, A.A. (1948). "An examination and further development of a formula arising in the problem of comparing two mean values", Biometrika 35, pp 88-96.

Aspin, A.A. (1949). "Tables for use in comparisons whose accuracy involves two variances, separately estimated", Biometrika 36, pp 290-302.

Fisher, R.A. and Yates, F. (1963). "Statistical tables for biological, agricultural and medical research", Edinburgh: Oliver and Boyd.

Lehmann, E.L. (1959). "Testing statistical hypotheses", John Wiley & Sons, Inc.

Lehmann, E.L. (1975). "Nonparametrics: Statistical methods based on ranks", McGraw-Hill.

Lindley, D.V.(1965). "Introduction to probability and statistics from a Bayesian viewpoint", Cambridge University Press.

Linnik, Yu.V. (1963). "On the Behrens-Fisher problem", Bull. Inst. Intern. Statist. 40 (2), pp 833-841.

Scheffé, H. (1943). "On solutions of the Behrens-Fisher problem, based on the t-distribution", Ann. of Math. Stat. 19, pp 35-44.

Scheffé, H. (1970). "Practical solutions of the Behrens-Fisher problem", J. Amer. Statist. Ass. 65, pp 1501-1508.

Sverdrup, E. (1967). "Laws and chance variations", Vol. II, North-Holland Publishing Company.

Wang, Y.Y. (1971). "Probabilities of the Type I errors of the Welch tests for the Behrens-Fisher problem", J. Amer. Statist. Ass. 66, pp 605-608.

Welch, B.L. (1937). "The significance of the difference between two means when the population variances are unequal", Biometrika 29, pp 350-362.

Welch, B.L. (1947). " The generalizations of Student's problem when several different population variances are involved", Biometrika 34, pp 28-35.

Yuen Fung, K. (1979). "A Monte Carlo study of the studentized Wilcoxon statistic for the Behrens-Fisher problem", J. Statist. Comput. Simul., 10, pp 15-24.

FIGURE LEGENDS

Fig. 1 The asymptotic significance levels of the T- and W-test for $\nu = 1$, $1/2$ and $1/3$, where $\nu = \lim(m/n)$ as $m,n \to \infty$ and $\theta = \sigma^2/\tau^2$.

Fig. 2 The significance levels of the T-est for $m = n = 6$ and 10 and $m = 5$, $n = 15$ compared with the asymptotic significance levels for $\nu = 1$ and $1/3$.

Fig. 3 The significance levels of the T-,U- and V-tests for (a) $m=n=10$ and (b) $m=5$, $n=15$.

Fig. 4 The power functions of the T-, U- and V-tests for $m=5$, $n=15$ and (a) $\theta = 1$, (b) $\theta = 4$ and (c) $\theta = 1/4$. The non-centrality parameter is $\delta = (\xi-\eta)/\sqrt{\sigma^2/m+\tau^2/n}$.

Fig. 5 The significance levels of the adaptive T-, U- and V-tests estimated by stochastic simulation (3000 samples) for (a) $m=n=10$, (b) $m=5$, $n=15$.

Fig. 6 The power functions of the adaptive T-, U- and V-tests for $m=5$, $n=15$ and (a) $\theta = 1$, (b) $\theta = 4$ and (c) $\theta = 1/4$. The non-centrality parameter is $\delta = (\xi-\eta)/\sqrt{\sigma^2/m+\tau^2/n}$.

Fig. 1

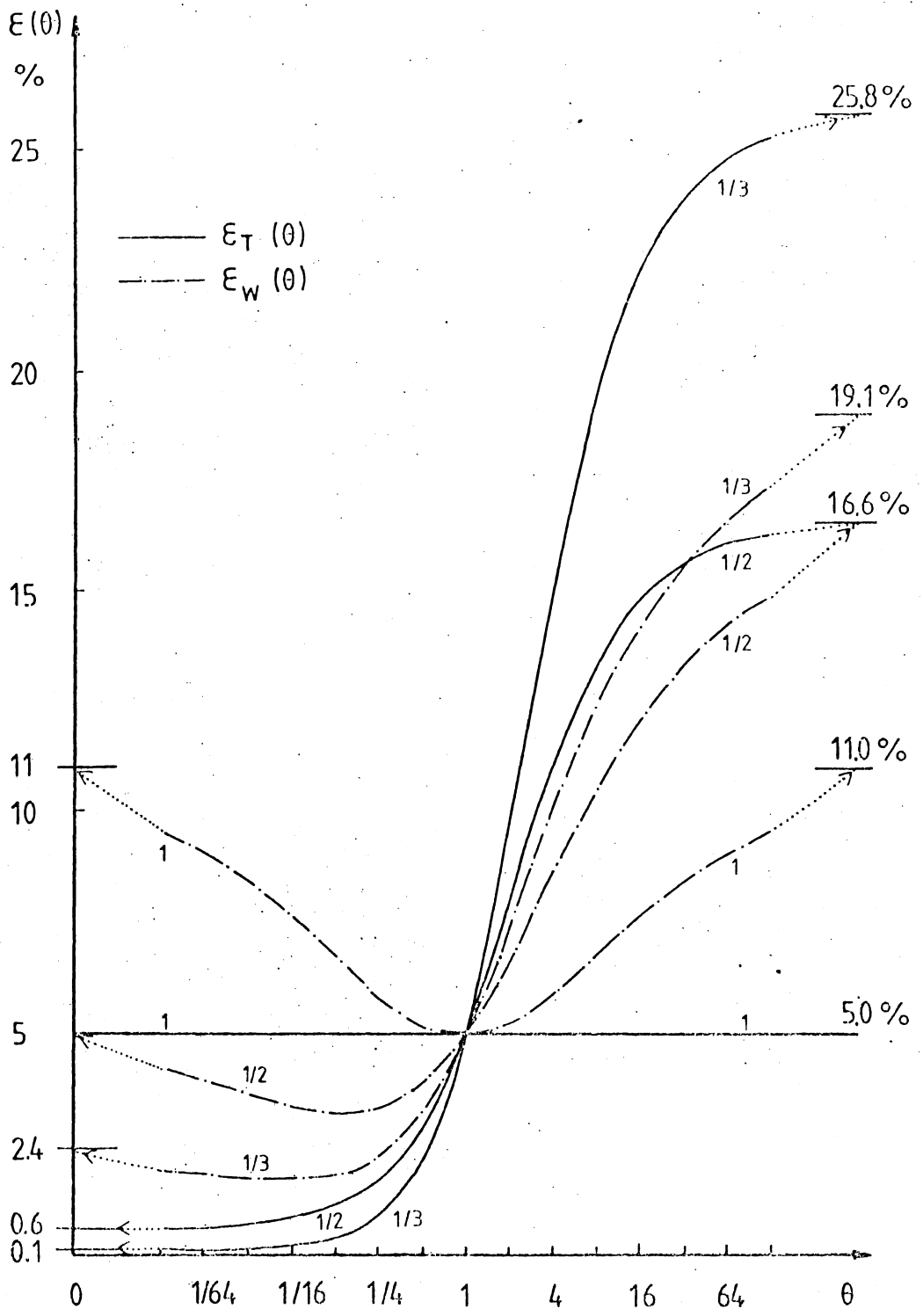

ε(θ)
%
25
20
15
11
10
5
2.4
0.6
0.1
ε_T (θ)
ε_W (θ)
25.8 %
19.1 %
16.6 %
11.0 %
5.0 %
1/3
1/3
1/2
1/2
1
1
1
1/2
1/3
1/2
1/3
0
1/64
1/16
1/4
1
4
16
64
θ

Fig. 2

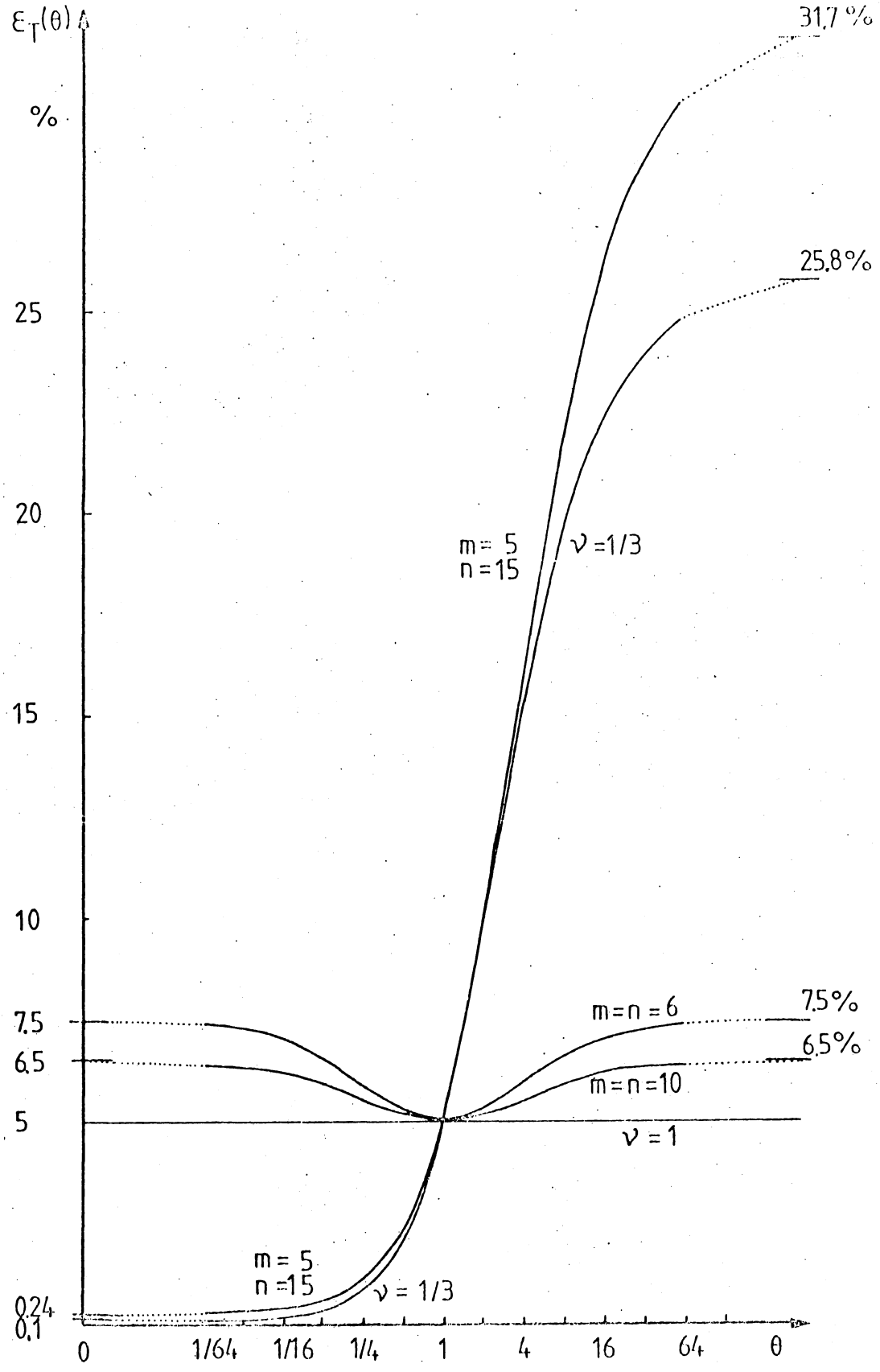
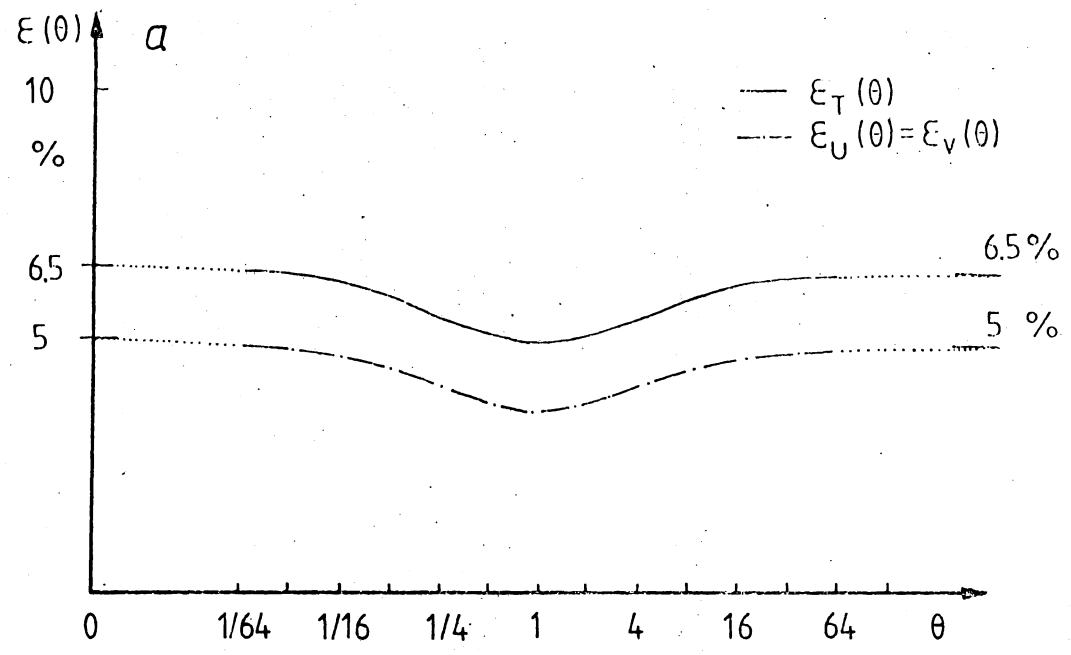

$\varepsilon_T(\theta)$ %
31.7 %
25.8%
25
20
15
10
7.5
6.5
5
0.24
0.1
m= 5
n=15
ν=1/3
m=n=6
7.5%
6.5%
m=n=10
ν=1
m = 5
n=15
ν = 1/3
0
1/64
1/16
1/4
1
4
16
64
θ

Fig. 3


ε(θ)
a
10
%
6.5
5
$\varepsilon_T(\theta)$
$\varepsilon_U(\theta)=\varepsilon_V(\theta)$
6.5 %
5 %
0
1/64
1/16
1/4
1
4
16
64
θ

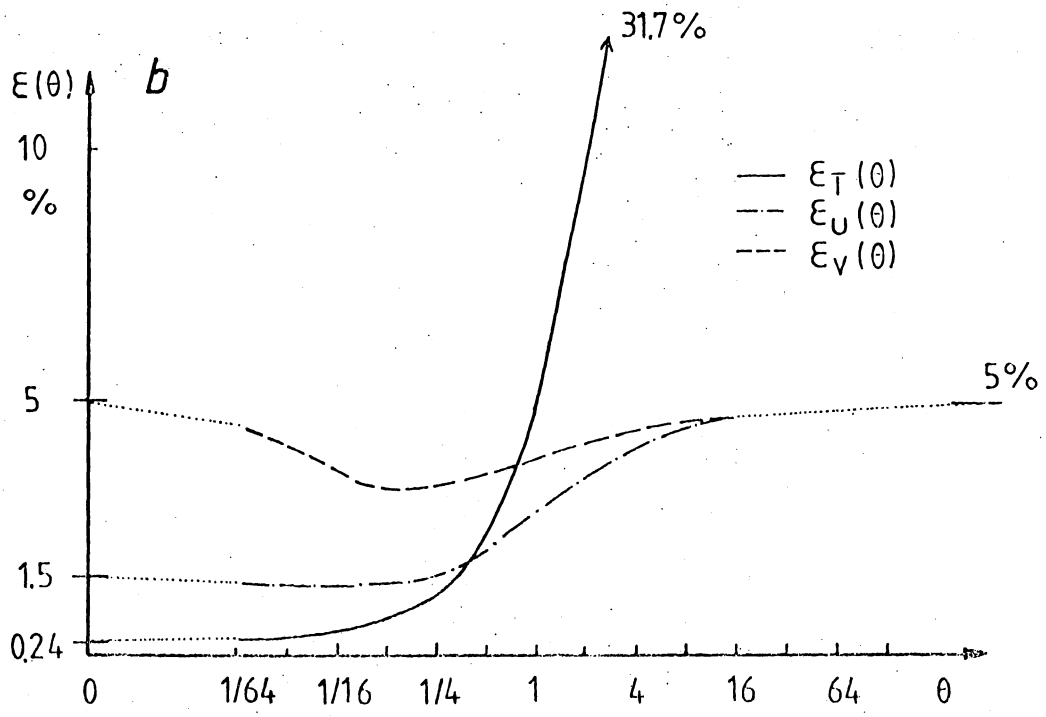

31.7 %
ε(θ)
b
10
%
5
1.5
0.24
$\varepsilon_T(\theta)$
$\varepsilon_U(\theta)$
$\varepsilon_V(\theta)$
5%
0
1/64
1/16
1/4
1
4
16
64
θ

Fig. 4

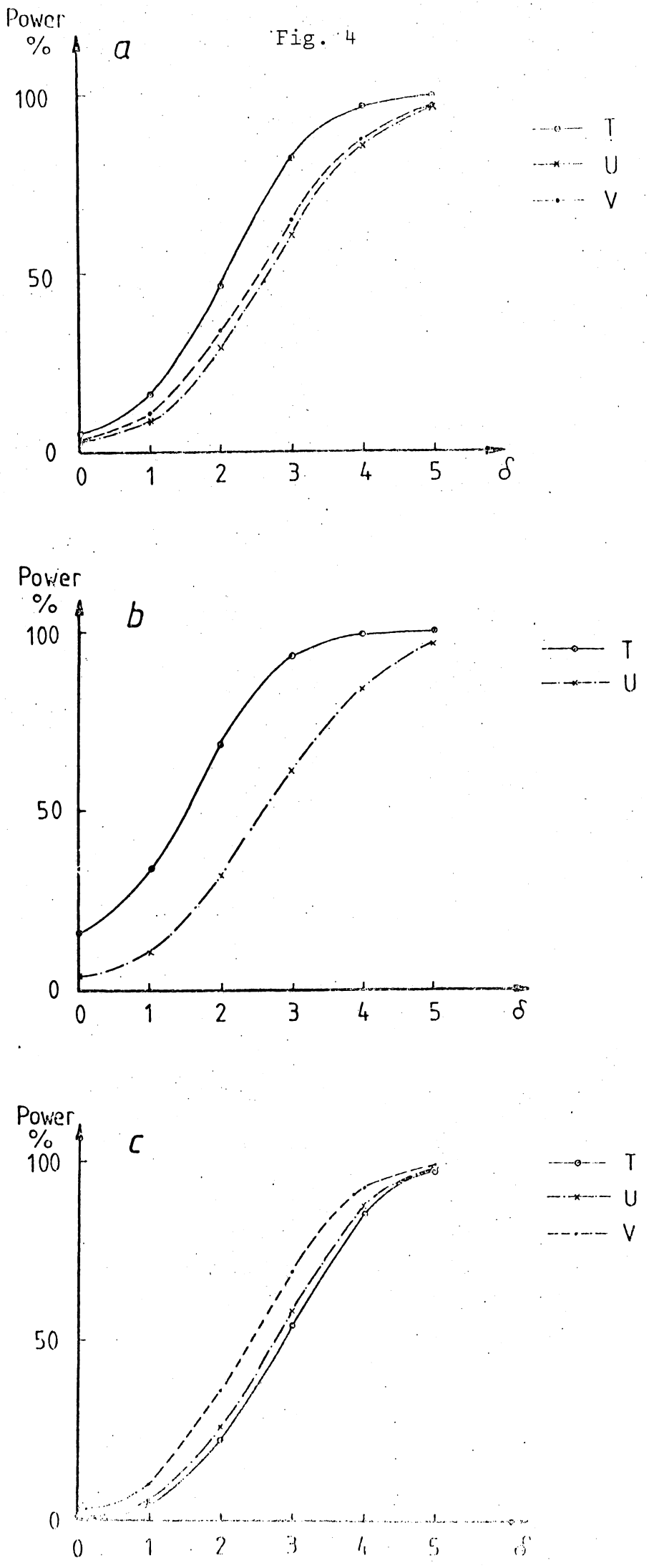

Power
%
a
100
50
0
0
1
2
3
4
5
δ
T
U
V
Power
%
b
100
50
0
0
1
2
3
4
5
δ
T
U
Power
%
c
100
50
0
0
1
2
3
4
5
δ
T
U
V

Fig. 5

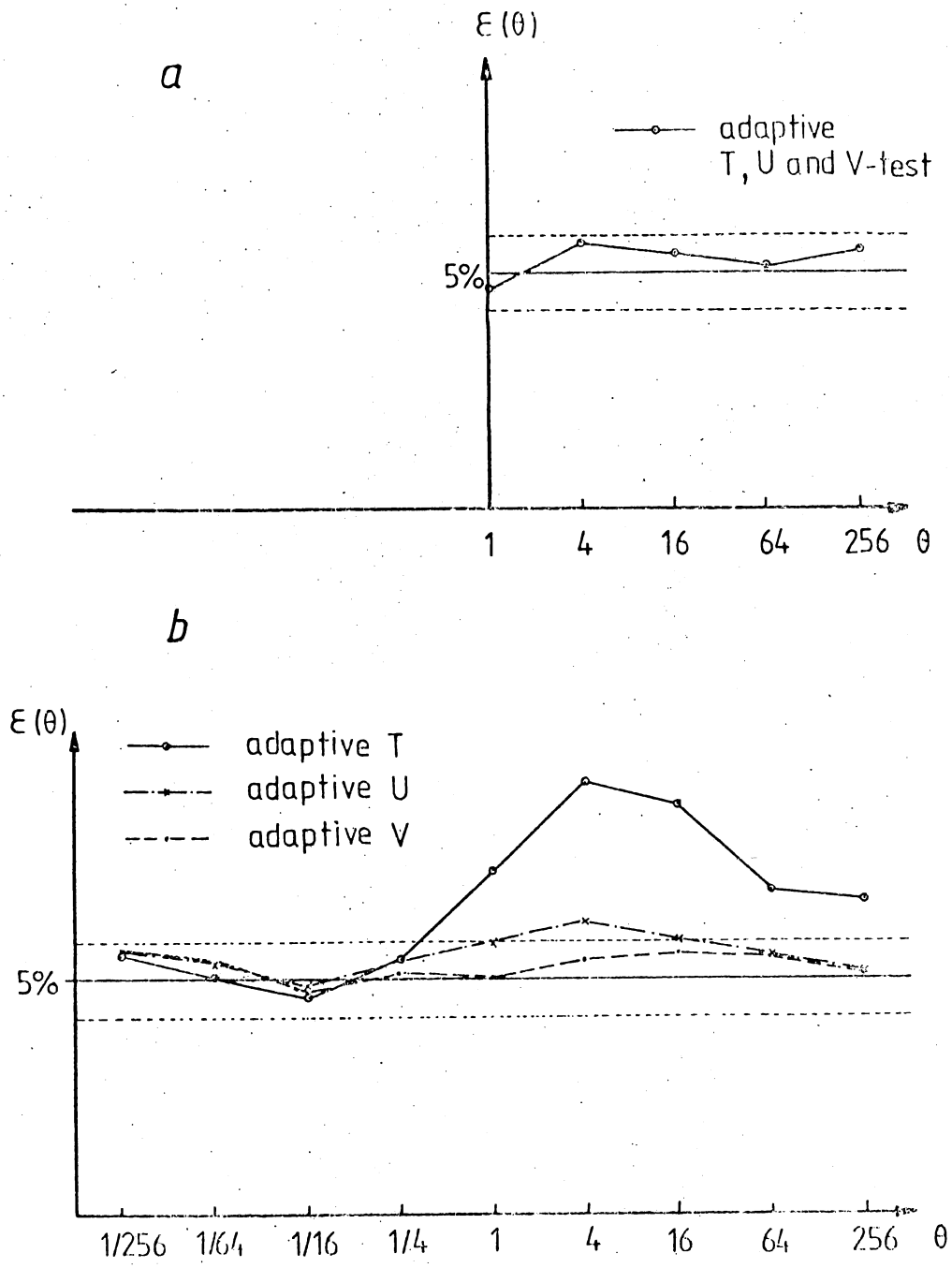
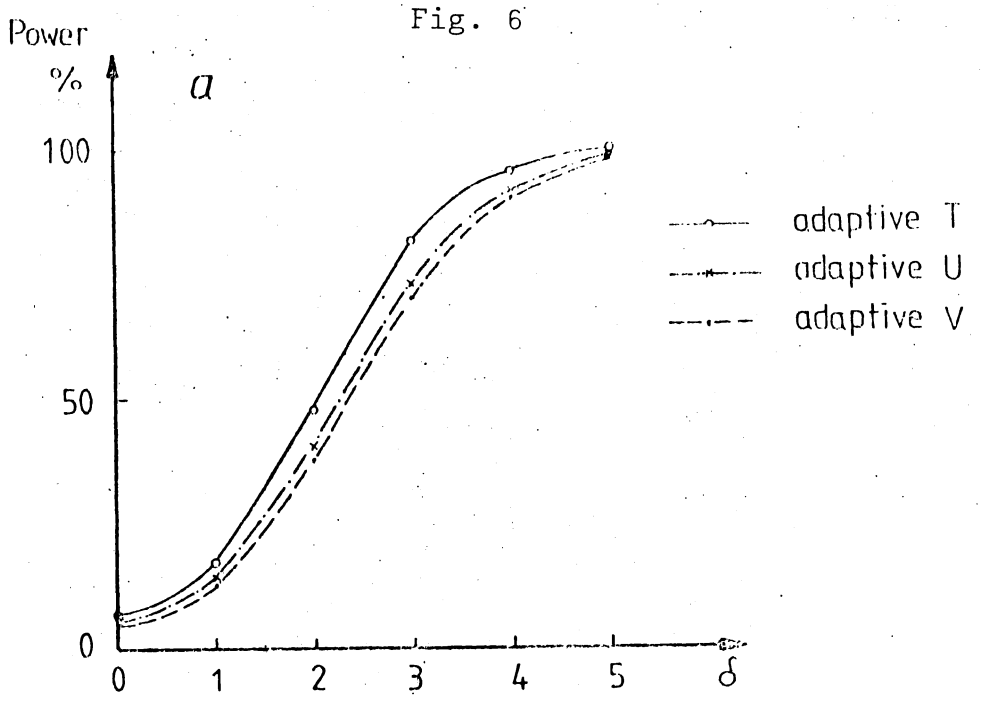
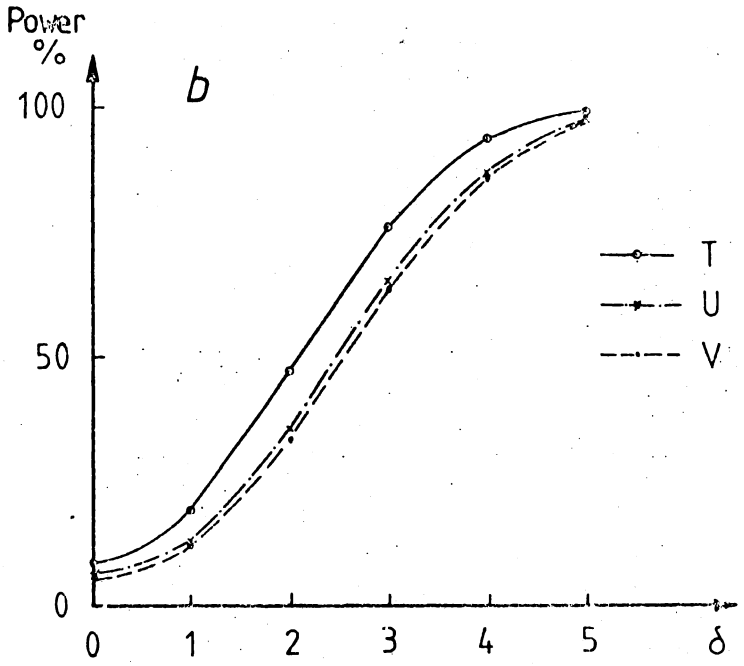
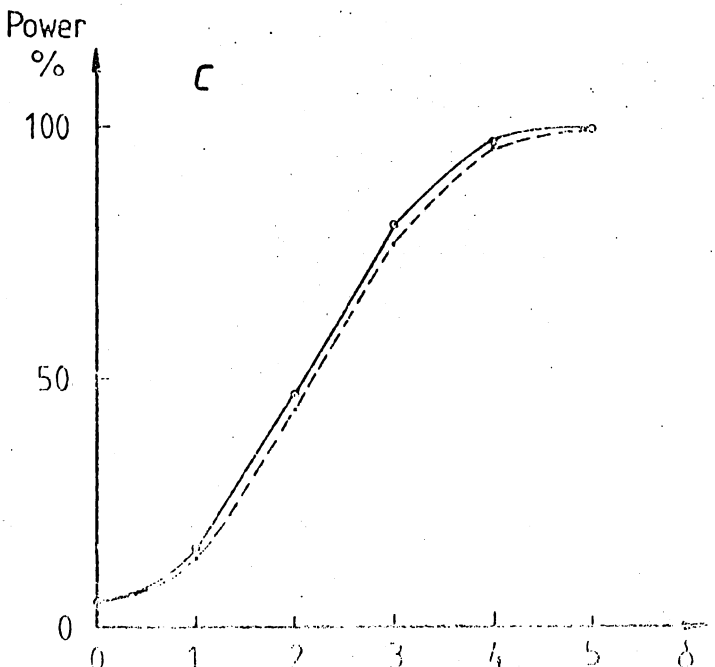

a
ε(θ)
adaptive
T, U and V-test
5%
1
4
16
64
256
θ
b
ε(θ)
adaptive T
adaptive U
adaptive V
5%
1/256
1/64
1/16
1/4
1
4
16
64
256
θ

Fig. 6


Power
%
a
100
50
0
0
1
2
3
4
5
δ
adaptive T
adaptive U
adaptive V


Power
%
b
100
50
0
0
1
2
3
4
5
δ
T
U
V


Power
%
c
100
50
0
0
1
2
3
4
5
δ